Lista 3

1. O arquivo de dados airfares.txt tem dados do preço("Fare") da passagem aérea só de ida(em dólares), e à distância("Distance") da cidade A a 17 outras cidades (em milhas) nos EUA. O interesse é modelar a passagem aérea em função da distância. O primeiro ajuste do modelo foi

$$Fare = \beta_0 + \beta_1 Distance + \epsilon$$

   Baseados na saída do ajuste do modelo, o analista de negocios conclui o siguinte:

   O coeficiente de regressão da variável preditora "Distância"é altamente significativa e o modelo explica 99,4% da variabilidade na variável Y , "Preço". Assim, o modelo é altamente eficaz tanto para a compreensão dos efeitos da distância no Preço, e para a previsão de valores futuros do Preço dado o valor da variável preditora, Distância.

   O que há de errado nessa conclusão? Será que o modelo de regressão simples parece ajustar os dados? Se não, descrever brevemente como o modelo pode ser melhorado.

2. Tem-se um conjunto de dados de preços de venda (em milhares de dólares) de casas aleatóriamente amostrados em Albuquerque, Novo México, na Primavera de 1993, juntamente com o tamanho da casa, por metro quadrado. O corretor de imóveis utilizou dados como estes para determinar o preço de venda adequado para uma casa. O arquivo dos dados é houseprices.txt.

   (a) Ajustar um modelo de regressão simples para a variável "price"com base em "sqft". Você observa evidências estatísticamente significativas na relação linear entre elas?.

   (b) Existem outliers, pontos de alavanca, ou pontos "ruins"de alavanca?.

   (c) Remova os pontos de alavanca "ruins"encontrados no item (b), e reajuste o modelo de regressão linear simples para "price"baseado em "sqft". Estes são aoutliers, pontos de alavanca, ou pontos de alavanca "ruins"?.

   (d) Utilize o conjunto de dados originais e calcule as distâncias do Cook utlizando a rutina no R *cooks.distance* e determine os pontos de alavanca "ruins". São esses pontos de alavanca "ruins"diferentes daqueles encontrados no item (b)?

3. Considere um exemplo real que envolva o gerenciamento de um porto no Canadá sobre Grandes Lagoas, onde deseja-se estimar a relação entre o volume de carga de um navio e do tempo necessário para carregar e descarregar essa carga. Essa relação será utilizado para fins de planejamento, bem como para fazer comparações com a produtividade de outros portos. Tem-se registros de toneladas("Tonnage") carregadas e descarregadas, como também, o tempo("Time") utilizado no porto pelas 31 embarcações

de líquido que foram utilizados no porto durante o último verão. Esses dados encontra-se no arquivo glakes.txt. Considere os seguintes modelos:

$$\begin{aligned} \text{Modelo 1:} Time &= \beta_0 + \beta_1 Tonnage + \epsilon \\ \text{Modelo 2:} Time &= \beta_0 + \beta_1 Tonnage^{0.25} + \epsilon \\ \text{Modelo 3:} log(Time) &= \beta_0 + \beta_1 Tonnage + \epsilon \end{aligned}$$

Qual desses modelos fornece melhor ajuste?